*Biometrika* (2016), **103**, 3, *pp.* 734–741
*Printed in Great Britain*
doi: 10.1093/biomet/asw031
Advance Access publication 25 July 2016

# A goodness-of-fit test for structural nested mean models

By S. Yang and J. J. Lok

*Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, Massachusetts 02115, U.S.A.*

shuyang@hsph.harvard.edu jlok@hsph.harvard.edu

Summary

Coarse structural nested mean models are tools for estimating treatment effects from longitudinal observational data with time-dependent confounding. There is, however, no guidance on how to specify the treatment effect model, and model misspecification can lead to bias. We derive a goodness-of-fit test based on modified over-identification restrictions tests for evaluating a treatment effect model, and show that our test is doubly robust in the sense that, with a correct treatment effect model, the test has the correct Type I error if either the treatment initiation model or a nuisance regression outcome model is correctly specified. In a simulation study, we show that the test has correct Type I error and can detect model misspecification. We use the test to study how the timing of antiretroviral treatment initiation after HIV infection predicts the effect of one year of treatment in HIV-positive patients with acute and early infection.

*Some key words*: Causal inference; Estimating equation; HIV/AIDs; Over-identification restrictions test.

## 1. Introduction

In observational studies, there is often time-dependent confounding: some covariates are predictors of both the treatment and the outcome. These covariates may also be affected by the treatment history. In such cases, standard regression methods adjusting for the covariate history can lead to bias (Robins et al., 1992; Robins, 2000; Robins et al., 2000). Coarse structural nested mean models (Robins, 1998) are useful for handling time-varying confounding, but they depend on correct specification of the treatment effect model.

In this paper we propose a goodness-of-fit test for correct specification of the treatment effect model. The key insight is that a correctly specified treatment effect model leads to a larger number of unbiased estimating equations than parameters, which results in over-identification of the latter. Over-identification restrictions tests, also called Sargan tests or $J$-tests (Sargan, 1958; Hansen, 1982), are widely used in econometrics. The standard over-identification restrictions test, given by the minimized value of the generalized method of moments (Newey & McFadden, 1994) criterion function, has a $\chi^2$ limiting distribution, with degrees of freedom equal to the number of over-identification restrictions. In most situations, the minimum of the generalized method of moments criterion is obtained via a continuous iterative procedure that updates the parameter estimates until convergence (Hansen et al., 1996). Arellano & Bond (1991) showed that the test statistic based on one-step estimates other than the optimal generalized method of moments estimates is not robust and tends to over-reject even in large samples. In the statistics literature, generalized method of moments inference has been used as part of the quadratic inference function approach developed by Qu et al. (2000) and Lindsay & Qu (2003).

Coarse structural nested mean models result in an infinite number of estimating functions, indexed by a set of arbitrary functions $q$. The precision of the estimator depends on the estimating functions (Lok & DeGruttola, 2012). Generalized method of moments approaches provide optimal combinations of the parameter-identification estimating functions and the goodness-of-fit estimating functions. However, it is not clear which estimating functions should be used. Semiparametric efficiency theory allows us to derive an optimal set of estimating equations whose corresponding $z$-estimator achieves the semiparametric

efficiency bound (Robins, 1994). Combining optimal estimating equations with additional goodness-of-fit estimating equations allows simultaneous estimation and testing, as in the traditional over-identification approach, but it can unnecessarily increase estimation variability (Lindsay & Qu, 2003). The purpose of this paper is to introduce a different strategy that separates estimation and testing, so that the estimator attains optimality under the null model and the test has high power. To achieve this, we obtain parameter estimates by solving the optimal estimating equations with the number of equations equal to the number of parameters, rather than by minimizing an objective function. The over-identified restrictions, used only for testing, can be developed from some parametric specification of alternative models. Simulation studies show that our test statistic has correct size for large samples and high power in all the scenarios considered. Another advantage of the over-identification restrictions test is that no bootstrap is needed to compute the test statistic, which is valuable when working with large samples.

## 2. Coarse structural nested mean model analysis

We assume that $n$ subjects are monitored at discrete times $k = 0, \ldots, K+1$. Let $Y_k$ be the outcome at time $k$, and let $L_k$ be a vector of covariates at time $k$. Let $A_k$ be the treatment indicator, which equals 1 if the subject was on treatment at time $k$ and 0 otherwise. We use overbars to denote a variable's history; for example, $\bar{A}_k = \{A_t : t = 0, \ldots, k\}$ is the treatment information at times $0, \ldots, k$. We assume that once treatment is started, it is never discontinued, so each treatment regime corresponds to a treatment starting time and vice versa. Let $T$ be the actual treatment starting time, with $T = \infty$ if the subject never started the treatment during follow-up. We assume that the subjects constitute an independent sample from a larger population (Rubin, 1978), and for notational simplicity we drop the subscript $i$ for subjects. Up to § 4 we shall assume that all subjects are followed until time $K+1$. Let $Y_k^{(m)}$ be the outcome at time $k$, possibly counterfactual, had the subject started the treatment at time $m$. We assume that the subject's observed outcome $Y_k$ is equal to the potential outcome $Y_k^{(m)}$ if $m$ is the actual treatment starting time $T$; that is, $Y_k = Y_k^{(T)}$. We also assume that there is no unmeasured confounding (Robins et al., 1992); that is, for $0 \leqslant m \leqslant k \leqslant K+1$, $Y_k^{(\infty)}$ is conditionally independent of $A_m$ given $\bar{L}_m$ and $\bar{A}_{m-1}$. This assumption holds if $\bar{L}_m$ contains all prognostic factors for $Y_k^{(\infty)}$ that affect the treatment decision at time $m$. Finally, $X = (\bar{A}_K, \bar{L}_K, \bar{Y}_{K+1})$ denotes the full information on a subject. Let $P$ be the probability measure induced by $X$ and $P_n$ the empirical measure induced by $X_1, \ldots, X_n$, and define $P_n f(X) = n^{-1} \sum_{i=1}^n f(X_i)$.

Following Lok & DeGruttola (2012), we model the treatment effect as

$$\gamma_{m,\psi}^k(\bar{l}_m) = E(Y_k^{(m)} - Y_k^{(\infty)} \mid \bar{L}_m = \bar{l}_m, T = m;\ \psi) \quad (0 \leqslant m \leqslant k \leqslant K+1), \tag{1}$$

where $\psi \in \mathbb{R}^p$ with $p \in \mathbb{N}$ fixed. Model (1) compares the average potential outcomes under treatment starting at time $m$ and treatment never started, among the subgroup of subjects with covariate history $\bar{l}_m$ and $T = m$. As such, it constitutes a conditional causal effect. In observational studies, the treatment assignment mechanism is unknown. We model the probability of treatment initiation at time $m$ conditional on the past history as $p_\theta(m) = \mathrm{pr}(A_m = 1 \mid \bar{A}_{m-1} = \bar{0}, \bar{L}_m;\ \theta)$, where the dimension of $\theta$ is finite and fixed. Let $J_{\mathrm{trt}(\theta)}(X)$ denote the estimating function for $\theta_0$. Define $H_\psi(k) = Y_k - \gamma_{T,\psi}^k(\bar{L}_T)$. As proved in Robins et al. (1992),

$$E\{H_\psi(k) \mid \bar{L}_m, \bar{A}_{m-1} = \bar{0}, A_m\} = E\{Y_k^{(\infty)} \mid \bar{L}_m, \bar{A}_{m-1} = \bar{0}, A_m\} \quad (0 \leqslant m \leqslant k \leqslant K+1)$$

and therefore, by the assumption of no unmeasured confounding, $E\{H_\psi(k) \mid \bar{L}_m, \bar{A}_{m-1} = \bar{0}, A_m\} = E\{H_\psi(k) \mid \bar{L}_m, \bar{A}_{m-1} = \bar{0}\}$. For any measurable, bounded function $q_m^k : \bar{\mathcal{L}}_m \to \mathbb{R}^p$ $(m = 0, \ldots, K)$, let

$$G_{(\psi,\theta,q)}(X) = \sum_{m=0}^{K} \sum_{k=m+1}^{K+1} q_m^k(\bar{L}_m)\big[H_\psi(k) - E\{H_\psi(k) \mid \bar{L}_m, \bar{A}_{m-1} = \bar{0}\}\big]\{A_m - p_\theta(m)\}. \tag{2}$$

Then $E\{G_{(\psi,\theta,q)}(X)\} = E[E\{G_{(\psi,\theta,q)}(X) \mid \bar{L}_m, \bar{A}_m\}] = 0$ (Lok & DeGruttola, 2012). Therefore, $P_n\{G_{(\psi,\theta,q)}(X)^{\mathrm{T}} J_{\mathrm{trt}(\theta)}(X)^{\mathrm{T}}\}^{\mathrm{T}} = 0$ are stacked unbiased estimating equations for both the parameter

$\psi$ and the nuisance parameter $\theta$. For simplicity, we will suppress the dependence of the estimating functions on $X$; for example, $P_n G_{(\psi,\theta,q)}$ is shorthand for $P_n G_{(\psi,\theta,q)}(X)$. Sometimes we will also drop the dependence on the parameters.

To derive the optimal estimating equation, and hence the optimal estimator, we assume that for $m = 0, \ldots, K$ and $k, s$ with $m+1 \leqslant k, s \leqslant K+1$, $\mathrm{cov}\{H(k), H(s) \mid \bar{L}_m, \bar{A}_{m-1} = \bar{0}, A_m\}$ does not depend on $A_m$. This is a working assumption only, which allows us to derive a closed-form solution for the optimal $q$ (Robins, 1994):

$$\begin{aligned}\begin{pmatrix} (q_m^{m+1,\mathrm{opt}})^{\mathrm{T}} \\ \vdots \\ (q_m^{K+1,\mathrm{opt}})^{\mathrm{T}} \end{pmatrix} = &\left[ \mathrm{var}\left\{ \begin{pmatrix} H_\psi(m+1) \\ \vdots \\ H_\psi(K+1) \end{pmatrix} \middle| \bar{L}_m, \bar{A}_{m-1} = \bar{0} \right\} \right]^{-1} \\ &\times \left[ E\left\{ \frac{\partial}{\partial \psi^{\mathrm{T}}} \begin{pmatrix} H_\psi(m+1) \\ \vdots \\ H_\psi(K+1) \end{pmatrix} \middle| \bar{L}_m, \bar{A}_{m-1} = \bar{0}, A_m = 1 \right\} \right. \\ &\left. - E\left\{ \frac{\partial}{\partial \psi^{\mathrm{T}}} \begin{pmatrix} H_\psi(m+1) \\ \vdots \\ H_\psi(K+1) \end{pmatrix} \middle| \bar{L}_m, \bar{A}_m = \bar{0} \right\} \right]. \end{aligned} \tag{3}$$

*Remark* 1. The optimal vector $q^{\mathrm{opt}}$ depends on the unknown $\psi$ and the true distribution through conditional expectations. Following Lok & DeGruttola (2012), we use a preliminary consistent estimate $\hat{\psi}_{\mathrm{p}}$ to replace $\psi$ in $E\{H_\psi(k) \mid \bar{L}_m, \bar{A}_{m-1} = \bar{0}\}$ and $q^{\mathrm{opt}}$. Also, we replace the unknown conditional expectations by estimators under working models, and write $E_{\xi_1}\{H_{\psi_{\mathrm{p}}}(k) \mid \bar{L}_m, \bar{A}_{m-1} = \bar{0}\}$ and $q^{k,\mathrm{opt}}_{m,\psi_{\mathrm{p}},\xi_1,\xi_2}$ to reflect their dependence on nuisance parameters $\xi_1$, $\xi_2$ and $\psi_{\mathrm{p}}$. Denote the estimating functions for $\xi_1$, $\xi_2$ and $\psi_{\mathrm{p}}$ by $J_{1(\xi_1,\psi_{\mathrm{p}})}$, $J_{2(\xi_2)}$ and $G_{\mathrm{p}(\psi_{\mathrm{p}},\xi_2)}$. By construction, the dimension of $q_m^{k,\mathrm{opt}}$ is $p$, so the estimating function (2) with (3) has the same dimension as $\psi$. Under certain modelling assumptions and regularity conditions for the estimating functions (see van der Vaart, 2000, §§ 5.2 and 5.3), the resulting $z$-estimator is consistent and asymptotically normal. Technical details are given in the Supplementary Material. Specifically, $\gamma^k_{m,\psi}$ must be correctly specified. In contrast, the estimator remains consistent for $\psi$ if either $E_{\xi_1}\{H_\psi(k) \mid \bar{L}_m, \bar{A}_{m-1} = \bar{0}\}$ or $p_\theta(m)$ is correctly specified. Thus, the estimator is doubly robust (Robins & Rotnitzky, 2001; van der Laan & Robins, 2003).

## 3. GOODNESS-OF-FIT TEST

Misspecification of the treatment effect model causes bias in treatment effect estimation. Here we develop tests for specification of the treatment effect model based on over-identification restrictions tests. For a correctly specified model, a new set of unbiased estimating functions which differ from the optimal ones used for estimation should be close to zero when evaluated at the optimal estimator. This is formalized in the following theorem.

THEOREM 1 (Goodness-of-fit test). *Let the treatment effect model be $\gamma^k_{m,\psi}(\bar{l}_m)$ and let $H_\psi(k) = Y_k - \gamma^k_{T,\psi}(\bar{l}_T)$. Choose a set of functions $\{\tilde{q}^k_m(\bar{L}_m) \in \mathbb{R}^\nu : 0 \leqslant m < k \leqslant K+1\}$, with $\nu$ a finite and fixed number, which are different from the optimal choice $q^{k,\mathrm{opt}}_m$. Let*

$$\tilde{G}_{(\psi,\psi_{\mathrm{p}},\xi,\theta)} = \sum_{m=0}^{K} \sum_{k=m+1}^{K+1} \tilde{q}^k_{m,\xi_2}(\bar{L}_m)\big[H_\psi(k) - E_{\xi_1}\{H_{\psi_{\mathrm{p}}}(k) \mid \bar{L}_m, \bar{A}_{m-1} = \bar{0}\}\big]\{A_m - p_\theta(m)\}. \tag{4}$$

*The null hypothesis $H_0$ is that $\gamma^k_m(\bar{l}_m)$ is correctly specified and either $E_{\xi_1}\{H_\psi(k) \mid \bar{L}_m, \bar{A}_{m-1} = \bar{0}\}$ or $p_\theta(m)$ is correctly specified. Under $H_0$ and all the required regularity conditions for estimating functions in van der Vaart (2000 §§ 5.2 and 5.3), the goodness-of-fit test statistic* $\mathrm{GOF} = n\{P_n \tilde{G}_{(\hat{\psi},\hat{\psi}_{\mathrm{p}},\hat{\xi},\hat{\theta})}\}^{\mathrm{T}} \hat{\Sigma}^{-1} P_n \tilde{G}_{(\hat{\psi},\hat{\psi}_{\mathrm{p}},\hat{\xi},\hat{\theta})}$

*tends to* $\chi^2(\nu)$ *in distribution as* $n \to \infty$, *where* $\Sigma$ *is the covariance matrix of* $\Phi_{(\psi_0,\psi_0,\xi_0,\theta_0)}$, *with*

$$\Phi = -\begin{pmatrix} P\dfrac{\partial}{\partial\psi}\tilde{G} \\ P\dfrac{\partial}{\partial\psi_{\mathrm{p}}}\tilde{G} \\ P\dfrac{\partial}{\partial\xi_1}\tilde{G} \\ 0 \\ P\dfrac{\partial}{\partial\theta}\tilde{G} \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} P\dfrac{\partial}{\partial\psi}G^* & P\dfrac{\partial}{\partial\psi_{\mathrm{p}}}G^* & P\dfrac{\partial}{\partial\xi_1}G^* & 0 & P\dfrac{\partial}{\partial\theta}G^* \\ 0 & P\dfrac{\partial}{\partial\psi_{\mathrm{p}}}G_{\mathrm{p}} & 0 & P\dfrac{\partial}{\partial\xi_2}G_{\mathrm{p}} & 0 \\ 0 & P\dfrac{\partial}{\partial\psi_{\mathrm{p}}}J_1 & P\dfrac{\partial}{\partial\xi_1}J_1 & 0 & 0 \\ 0 & 0 & 0 & P\dfrac{\partial}{\partial\xi_2}J_2 & 0 \\ 0 & 0 & 0 & 0 & P\dfrac{\partial}{\partial\theta}J_{\mathrm{trt}} \end{pmatrix}^{-1} \begin{pmatrix} G^* \\ G_{\mathrm{p}} \\ J_1 \\ J_2 \\ J_{\mathrm{trt}} \end{pmatrix},$$

*where* $G^*$ *is the estimating function* (2) *with* (3), $G_{\mathrm{p}}$, $J_1$, $J_2$ *and* $J_{\mathrm{trt}}$ *are as defined in Remark* 1, *and* $\hat{\Sigma}$ *is the estimated variance of* $\{\hat{\Phi}_{(\hat{\psi},\hat{\psi}_{\mathrm{p}},\hat{\xi},\hat{\theta})}(X_i) : i = 1, \ldots, n\}$, *with* $\hat{\Phi}$ *obtained by replacing* $P$ *in* $\Phi$ *with* $P_n$.

We state here the key steps of the proof; the details can be found in the Supplementary Material. We first establish the asymptotic distribution of $n^{1/2}P_n\tilde{G}_{(\hat{\psi},\hat{\psi}_{\mathrm{p}},\hat{\xi},\hat{\theta})}$. A key step is to linearize $n^{1/2}P_n\tilde{G}_{(\hat{\psi},\hat{\psi}_{\mathrm{p}},\hat{\xi},\hat{\theta})}$ as $n^{1/2}P_n\Phi_{(\psi_0,\psi_0,\xi_0,\theta_0)}$ for some function $\Phi$, and apply the central limit theorem. To do so, we assume that the functions $\tilde{G}_{(\psi,\psi_{\mathrm{p}},\xi,\theta)}$ form a Donsker class. Using Lemma 19.24 of van der Vaart (2000), we have $n^{1/2}(P_n - P)\tilde{G}_{(\hat{\psi},\hat{\psi}_{\mathrm{p}},\hat{\xi},\hat{\theta})} = n^{1/2}(P_n - P)\tilde{G}_{(\psi_0,\psi_0,\xi_0,\theta_0)} + o_{\mathrm{p}}(1)$. We can then express the asymptotic linear representation of $\tilde{G}_{(\hat{\psi},\hat{\psi}_{\mathrm{p}},\hat{\xi},\hat{\theta})}$ as $\Phi_{(\psi_0,\psi_0,\xi_0,\theta_0)}$, which is a linear combination of $G^*$, $\tilde{G}$, $G_{\mathrm{p}}$, $J_1$, $J_2$ and $J_{\mathrm{trt}}$, all evaluated at the truth.

The goodness-of-fit test statistic is doubly robust in the sense that for the $\chi^2$ limiting distribution to hold, we require only that either $E_{\xi_1}\{H_\psi(k) \mid \bar{L}_m, \bar{A}_{m-1} = \bar{0}\}$ or $p_\theta(m)$ be correctly specified. This property adds protection against possible misspecification of the nuisance models.

The goodness-of-fit test with an arbitrary $\tilde{q}$ may have low power. We propose the following procedure for choosing $\tilde{q}$. We specify two models: the null model $\gamma_\psi^*$ and an alternative model $\tilde{\gamma}_\psi$. We can derive $q^{*\mathrm{opt}}$ and $\tilde{q}^{\mathrm{opt}}$ as in (3) with $\gamma_\psi^*$ and $\tilde{\gamma}_\psi$, respectively. We use $q^{*\mathrm{opt}}$ in the parameter-identification estimating function (2) and $\tilde{q}^{\mathrm{opt}}$ in the goodness-of-fit estimating function (4). Our simulation study shows that the goodness-of-fit test with $\tilde{q}^{\mathrm{opt}}$ has high power in the scenarios considered.

## 4. Extension of goodness-of-fit test in the presence of censoring

We use inverse probability of censoring weighting (Robins et al., 1995; Hernán et al., 2005) to accommodate subjects lost to follow-up. Let $C_p = 0$ indicate that a subject remains in the study at time $p$. Following Lok & DeGruttola (2012), we assume that censoring is missing at random; that is, $(\bar{L}, \bar{A})$ is independent of $C_{k+1}$ given $\bar{L}_k, \bar{A}_k, \bar{C}_k = \bar{0}$. Then $\mathrm{pr}(A_m = 1 \mid \bar{L}_m, \bar{A}_{m-1} = \bar{0}, \bar{C}_m = \bar{0}) = \mathrm{pr}(A_m = 1 \mid \bar{L}_m, \bar{A}_{m-1} = \bar{0}) = p_\theta(m)$. Define the inverse probability of censoring weighted estimating functions $G^{\mathrm{c}*}$ and $\tilde{G}^{\mathrm{c}}$ using weights $W^k_{m,\eta} = \{\prod_{p=m+1}^{k} \mathrm{pr}(C_p = 0 \mid \bar{L}_{p-1}, \bar{A}_{p-1}, \bar{C}_{p-1} = \bar{0}; \eta)\}^{-1}$; see Lok & DeGruttola (2012). We assume that the censoring model is correctly specified and identified with estimating functions $J_{\mathrm{cen}(\eta)}$. Similarly, we denote the inverse probability of censoring weighted estimating function for the preliminary estimator $\hat{\psi}_{\mathrm{p}}$ by $G^{\mathrm{c}}_{\mathrm{p}}$. For the nuisance regression outcome models, the regression was also weighted. Define the goodness-of-fit test statistic in the presence of censoring by $\mathrm{GOF}^{\mathrm{c}} = n\{P_n\tilde{G}^{\mathrm{c}}_{(\hat{\psi},\hat{\psi}_{\mathrm{p}},\hat{\xi},\hat{\theta},\hat{\eta})}\}^{\mathrm{T}}(\hat{\Sigma}^{\mathrm{c}})^{-1}P_n\tilde{G}^{\mathrm{c}}_{(\hat{\psi},\hat{\psi}_{\mathrm{p}},\hat{\xi},\hat{\theta},\hat{\eta})}$, where $\hat{\Sigma}^{\mathrm{c}}$ is the estimated variance of

$\{\hat{\Phi}^{\mathrm{c}}_{(\hat{\psi},\hat{\psi}_{\mathrm{p}},\hat{\xi},\hat{\theta},\hat{\eta})}(X_i) : i = 1, \ldots, n\}$, with

$$\Phi^{\mathrm{c}} = \begin{pmatrix} P\frac{\partial}{\partial\psi}\tilde{G}^{\mathrm{c}} \\ P\frac{\partial}{\partial\psi_{\mathrm{p}}}\tilde{G}^{\mathrm{c}} \\ P\frac{\partial}{\partial\xi_1}\tilde{G}^{\mathrm{c}} \\ 0 \\ P\frac{\partial}{\partial\theta}\tilde{G}^{\mathrm{c}} \\ P\frac{\partial}{\partial\eta}\tilde{G}^{\mathrm{c}} \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} P\frac{\partial}{\partial\psi}G^{\mathrm{c}*} & P\frac{\partial}{\partial\psi_{\mathrm{p}}}G^{\mathrm{c}*} & P\frac{\partial}{\partial\xi_1}G^{\mathrm{c}*} & 0 & P\frac{\partial}{\partial\theta}G^{\mathrm{c}*} & P\frac{\partial}{\partial\eta}G^{\mathrm{c}*} \\ 0 & P\frac{\partial}{\partial\psi_{\mathrm{p}}}G^{\mathrm{c}}_{\mathrm{p}} & 0 & P\frac{\partial}{\partial\xi_2}G^{\mathrm{c}}_{\mathrm{p}} & 0 & P\frac{\partial}{\partial\eta}G^{\mathrm{c}}_{\mathrm{p}} \\ 0 & P\frac{\partial}{\partial\psi_{\mathrm{p}}}J_1 & P\frac{\partial}{\partial\xi_1}J_1 & 0 & 0 & P\frac{\partial}{\partial\eta}J_1 \\ 0 & 0 & 0 & P\frac{\partial}{\partial\xi_2}J_2 & 0 & P\frac{\partial}{\partial\eta}J_2 \\ 0 & 0 & 0 & 0 & P\frac{\partial}{\partial\theta}J_{\mathrm{trt}} & 0 \\ 0 & 0 & 0 & 0 & 0 & P\frac{\partial}{\partial\eta}J_{\mathrm{cen}} \end{pmatrix}^{-1} \begin{pmatrix} G^{\mathrm{c}*} \\ G^{\mathrm{c}}_{\mathrm{p}} \\ J_1 \\ J_2 \\ J_{\mathrm{trt}} \\ J_{\mathrm{cen}} \end{pmatrix}.$$

As proved in the Supplementary Material, subject to regularity conditions, GOF$^{\mathrm{c}}$ has an asymptotic $\chi^2$ distribution, with degrees of freedom equal to the dimension of $\tilde{G}^{\mathrm{c}}$.

## 5. SIMULATIONS

Our simulation study is based on the HIV data described in § 6. Following the approach described in a technical report available from the second author, we used an autoregressive model for the time course of the CD4 count $Y_k^{(\infty)}$ under no treatment, which may be more realistic in months $k = 6, \ldots, 30$ than at earlier times given the different behaviour of CD4 counts during the first six months after infection. Therefore, we simulated data in months 6 to 30. First, in each sample, two groups were simulated: 10% injection drug users and 90% non-injection drug users. The outcome was simulated as log-normal: $\log Y_6^{(\infty)} \sim N(6{\cdot}6, 0{\cdot}4^2)$ for injection drug users, and $N(6{\cdot}6, 0{\cdot}5^2)$ for non-injection drug users. For $k \geqslant 6$, $Y_{k+1}^{(\infty)} = -10 + Y_k^{(\infty)} + \epsilon_{k+1}$, where $\epsilon_k \sim N(0, \sigma_k^2)$, with $\sigma_k = 52{\cdot}375 - 1{\cdot}625k$ for $k = 7, \ldots, 19$ and $\sigma_k = 21{\cdot}5$ for $k = 20, \ldots, 30$. Second, $T$ was generated by $\mathrm{logit}\,\mathrm{pr}(T = m \mid T \geqslant m, \bar{L}_m) = -2{\cdot}4 - 0{\cdot}42\,(\mathrm{injdrug}) - 0{\cdot}0035 Y_m^{(\infty)} - 0{\cdot}026m$, where injdrug is an indicator of injection drug use. Finally, $Y_k = Y_k^{(\infty)} + \gamma_T^k(\bar{L}_T)$. We considered different models for $\gamma_m^k$. The censoring process was generated by $\mathrm{logit}\,\mathrm{pr}(C_{m+1} = 1 \mid \bar{C}_m = \bar{1}, \bar{L}_m) = 2 + 3\,(\mathrm{injdrug}) + 0{\cdot}1 Y_m^{1/2}$. Under this model, the average proportion of patients being censored before month 30 is about 42%.

The performance of the test statistics was assessed in terms of Type I error and power. We are interested in testing $H_0 : \gamma_{m,\psi}^k = (\psi_1 + \psi_2 m)(k-m)1_{(k>m)}$ versus $H_{\mathrm{a}} : \gamma_{m,\psi}^k \neq (\psi_1 + \psi_2 m)(k-m)1_{(k>m)}$. Under $H_{\mathrm{a}}$, we specified $\tilde{\gamma}$ for which the test should have optimal power. Four scenarios were considered. In scenarios (a) and (b), $\gamma$ is correctly specified; in scenarios (c) and (d), $\gamma$ is misspecified. In scenario (c), $\gamma$ is nested in $\tilde{\gamma}$; and in scenario (d), $\gamma$ is not nested in $\tilde{\gamma}$.

(a) True: $\gamma_{m,\psi}^k = (25 - 0{\cdot}7m)(k-m)$, $H_0 : \gamma_{m,\psi}^k = (\psi_1 + \psi_2 m)(k-m)$, and $\tilde{\gamma}_{m,\psi}^k = (\psi_1 + \psi_2 m + \psi_3 m^2)(k-m)$.

(b) True: $\gamma_{m,\psi}^k = (25 - 0{\cdot}7m)(k-m)$, $H_0 : \gamma_{m,\psi}^k = (\psi_1 + \psi_2 m + \psi_3 I(\mathrm{injdrug}))(k-m)$, and $\tilde{\gamma}_{m,\psi}^k = (\psi_1 + \psi_2 m + \psi_3 m^2)(k-m)$.

(c) True: $\gamma_{m,\psi}^k = (35 - 1{\cdot}1m + 0{\cdot}04m^2)(k-m)$, $H_0 : \gamma_{m,\psi}^k = (\psi_1 + \psi_2 m)(k-m)$, and $\tilde{\gamma}_{m,\psi}^k = (\psi_1 + \psi_2 m + \psi_3 m^2)(k-m)$.

(d) True: $\gamma_{m,\psi}^k = (25 - m + 0{\cdot}03m^2)(k-m)$, $H_0 : \gamma_{m,\psi}^k = (\psi_1 + \psi_2 m)(k-m)$, and $\tilde{\gamma}_{m,\psi}^k = (\psi_3 + \psi_4 m)(k-m)^{3/2}$.

We estimated the size and power by the frequency of rejecting $H_0$ in 1000 simulated datasets. We considered the following choices of $\tilde{q}$: (i) $\tilde{q}_m^k = 1$, a naive choice; (ii) $\tilde{q}_m^k = \tilde{\Delta}_m^k$, which is obtained from

Table 1. *Type I error estimates and power estimates* ($\times 100$) *for testing the null model* $H_0$ *by the proposed goodness-of-fit test statistic with* $\tilde{q} = 1$, $\tilde{\Delta}$ *or* $\tilde{q}^{\text{opt}}$ *and by the elaborated model fitting and testing approach over* 1000 *simulations under scenarios* (a)–(d)

| | Type I error estimates under scenario (a) | | | | Type I error estimates under scenario (b) | | | |
|---|---|---|---|---|---|---|---|---|
| | | GOF | | EMFT | | GOF | | EMFT |
| $n\backslash\tilde{q}$ | 1 | $\tilde{\Delta}_m^k$ | $\tilde{q}_m^{\text{opt},k}$ | | 1 | $\tilde{\Delta}_m^k$ | $\tilde{q}_m^{\text{opt},k}$ | |
| 500 | 5·4 | 4·4 | 5·1 | 4·8 | 8·1 | 9·1 | 10·3 | 12·2 |
| 1000 | 5·1 | 5·4 | 4·8 | 5·6 | 5·4 | 5·4 | 5·2 | 5·6 |
| 2000 | 4·7 | 5·1 | 4·9 | 5·4 | 4·8 | 5·2 | 4·9 | 5·3 |
| | Power estimates under scenario (c) | | | | Power estimates under scenario (d) | | | |
| | | GOF | | EMFT | | GOF | | EMFT |
| $n\backslash\tilde{q}$ | 1 | $\tilde{\Delta}_m^k$ | $\tilde{q}_m^{\text{opt},k}$ | | 1 | $\tilde{\Delta}_m^k$ | $\tilde{q}_m^{\text{opt},k}$ | |
| 500 | 10 | 20 | 56 | 52 | 13 | 26 | 48 | 25 |
| 1000 | 19 | 42 | 74 | 70 | 21 | 49 | 67 | 51 |
| 2000 | 41 | 73 | 90 | 90 | 39 | 73 | 88 | 73 |

GOF, the proposed goodness-of-fit test; EMFT, elaborated model fitting and testing approach.

formula (3) after replacing $\gamma$ by $\tilde{\gamma}$ and the covariance matrix by a working identity matrix; and (iii) $\tilde{q}_m^k = \tilde{q}_m^{k,\text{opt}}$, which is obtained from formula (3) upon replacing $\gamma$ by $\tilde{\gamma}$. The nuisance models are specified in the Supplementary Material. In addition to the goodness-of-fit test statistic, we considered an elaborated model fitting and testing approach, which combines the null model with $\tilde{\gamma}$, and tests whether the parameters corresponding to $\tilde{\gamma}$ are equal to zero. As can be seen from Table 1, the goodness-of-fit test procedure does not control Type I error well for scenario (b) with $n = 500$; however, in scenarios (a) and (b), it controls Type I error with all choices of $\tilde{q}$ for $n = 1000$ and $n = 2000$. This suggests that the $\chi^2$ distribution provides an accurate approximation to the finite-sample behaviour of the goodness-of-fit test statistic for moderate sample sizes. From scenarios (c) and (d), it can be seen that the goodness-of-fit test procedure with $\tilde{q}_m^{k,\text{opt}}$ has the highest power, and as the sample size increases, the power increases, confirming the theoretical results. The goodness-of-fit test procedure with $\tilde{q}_m^{\text{opt},k}$ is comparable to the elaborated model fitting and testing approach when testing nested models as in scenario (c); however, it shows more power in detecting nonnested models as in scenario (d), probably because the elaborated model fitting and testing approach fits a larger model and hence loses power.

## 6. Application

We used the proposed test to study how the timing of antiretroviral treatment initiation after HIV infection predicts the effect of one year of treatment in HIV-positive patients. We analysed data from the Acute Infection and Early Disease Research Program, which is a multicentre, observational cohort study of HIV-positive patients diagnosed during acute and early infection (Hecht et al., 2006). Dates of infection were estimated based on a stepwise algorithm that uses clinical and laboratory data (Smith et al., 2006). We included patients with CD4 and viral load measured within 12 months of the estimated date of infection, which resulted in 1696 patients. Let $Y_k$ be the patient's CD4 count at month $k$ after the estimated date of infection, and let $L_k$ be a vector of covariates including age, gender, race, injection drug use, CD4 count and viral load. We let $m$ range from 0 to 23 and $k$ from $\max(12, m + 1)$ to $\min(m + 12, 24)$, to avoid making extra modelling assumptions beyond those necessary to estimate the one-year treatment effect $\gamma_{m,\psi}^{m+12}$. We assumed that treatment can only be initiated at visit times. If $L_m$ was missing at a visit time, we carried the last observation forward. For intermediate missing outcomes, we imputed $Y_k$ by interpolation; 1·6% of the outcomes were missing just prior to onset of treatment, and in such cases we carried the last observation forward. The percentage of patients censored before month 24 is about 45·7%.

We started with a simple null model for the treatment effect, $H_0 : \gamma_{m,\psi}^k = (\psi_1 + \psi_2 m)(k - m)1_{(k>m)}$, and conducted three directed alternative-model tests directed at gender, age and injection drug use, as

Table 2. *Application of our proposed test to the HIV data: the optimal estimator fitting the null treatment effect model, showing point estimates (with 95% confidence intervals in parentheses), goodness-of-fit statistics, associated degrees of freedom, and p-values for the adequacy of the null model, by testing whether gender, age or injection drug use should be added into the model*

| $\hat{\psi}_1$ (95% CI) | $\hat{\psi}_2$ (95% CI) |
|---|---|
| 24·88 (21·61, 28·15) | −0·48 (−1·47, 0·52) |

| Goodness-of-fit test | Statistic | DF | $p$-value |
|---|---|---|---|
| Test directed at gender | 0·99 | 1 | 0·32 |
| Test directed at age | 0·80 | 1 | 0·37 |
| Test directed at injection drug use | 2·93 | 1 | 0·09 |

CI, confidence interval; DF, degrees of freedom.

suggested in the clinical literature. For the test directed at a certain variable $Z$, we calculated the goodness-of-fit test statistic with $\tilde{q}$ having the optimal form derived from $\tilde{\gamma}^k_{m,\psi} = (\psi_1 + \psi_2 m + \psi_3 Z)(k-m)1_{(k>m)}$. The nuisance models are specified in the Supplementary Material. Table 2 shows the results. The $p$-values are all greater than 0·05, which suggests that there is no significant evidence for rejection of the null model. The results indicate a benefit of antiretroviral treatment; for example, starting treatment at the estimated date of infection would lead to an expected added improvement in CD4 counts of $12\hat{\psi}_1 = 299$ cells/mm$^3$ after a year of therapy. Delaying treatment initiation may diminish the CD4 count gain associated with one year of treatment, since $\hat{\psi}_2 < 0$; however, this result is not statistically significant.

## ACKNOWLEDGEMENT

We thank James Robins, Victor DeGruttola, Susan Little and Davey Smith for insightful and fruitful discussions. This work was sponsored by the U.S. National Institutes of Health.

## SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online includes asymptotic properties of the optimal estimator and the goodness-of-fit test statistic, as well as details of the simulation and application.

## REFERENCES

ARELLANO, M. & BOND, S. (1991). Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Rev. Econ. Studies* **58**, 277–97.

HANSEN, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica* **50**, 1029–54.

HANSEN, L. P., HEATON, J. & YARON, A. (1996). Finite-sample properties of some alternative GMM estimators. *J. Bus. Econ. Statist.* **14**, 262–80.

HECHT, F. M., WANG, L., COLLIER, A., LITTLE, S., MARKOWITZ, M., MARGOLICK, J., KILBY, J. M., DAAR, E. & CONWAY, B. (2006). A multicenter observational study of the potential benefits of initiating combination antiretroviral therapy during acute HIV infection. *J. Infect. Dis.* **194**, 725–33.

HERNÁN, M. A., COLE, S. R., MARGOLICK, J., COHEN, M. & ROBINS, J. M. (2005). Structural accelerated failure time models for survival analysis in studies with time-varying treatments. *Pharmacoepidemiol. Drug Safety* **14**, 477–91.

LINDSAY, B. G. & QU, A. (2003). Inference functions and quadratic score tests. *Statist. Sci.* **18**, 394–410.

LOK, J. J. & DEGRUTTOLA, V. (2012). Impact of time to start treatment following infection with application to initiating HAART in HIV-positive patients. *Biometrics* **68**, 745–54.

NEWEY, W. K. & MCFADDEN, D. (1994). Large sample estimation and hypothesis testing. *Handbook of Econometrics* **4**, 2111–245.

QU, A., LINDSAY, B. G. & LI, B. (2000). Improving generalised estimating equations using quadratic inference functions. *Biometrika* **87**, 823–36.

Robins, J. M. (1994). Correcting for non-compliance in randomized trials using structural nested mean models. *Commun. Statist.* A **23**, 2379–412.

Robins, J. M. (1998). Correction for non-compliance in equivalence trials. *Statist. Med.* **17**, 269–302.

Robins, J. M. (2000). Marginal structural models versus structural nested models as tools for causal inference. In *Statistical Models in Epidemiology, the Environment, and Clinical Trials*. New York: Springer, pp. 95–133.

Robins, J. M., Blevins, D., Ritter, G. & Wulfsohn, M. (1992). G-estimation of the effect of prophylaxis therapy for Pneumocystis carinii pneumonia on the survival of AIDS patients. *Epidemiology* **3**, 319–36.

Robins, J. M., Hernan, M. A. & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology* **11**, 550–60.

Robins, J. M. & Rotnitzky, A. (2001). Comment on the Bickel and Kwon article 'Inference for semiparametric models: Some questions and an answer'. *Statist. Sinica* **11**, 863–960.

Robins, J. M., Rotnitzky, A. & Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J. Am. Statist. Assoc.* **90**, 106–21.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Ann. Statist.* **6**, 34–58.

Sargan, J. D. (1958). The estimation of economic relationships using instrumental variables. *Econometrica* **26**, 393–415.

Smith, D. M., Strain, M. C., Frost, S. D., Pillai, S. K., Wong, J. K., Wrin, T., Liu, Y., Petropolous, C. J., Daar, E. S. & Little, S. J. et al. (2006). Lack of neutralizing antibody response to HIV-1 predisposes to superinfection. *Virology* **355**, 1–5.

van der Laan, M. J. & Robins, J. M. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. Berlin: Springer.

van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge: Cambridge University Press.

[*Received December* 2015. *Revised June* 2016]